






It is All in the Wrist: Wearable Sleep Staging in a Clinical Population versus Reference Polysomnography

Bernice M Wulterkens ^{1,2,*}Pedro Fonseca ^{1,2,*}Lieke WA Hermans ¹Marco Ross ³Andreas Cerny ³Peter Anderer ³Xi Long ^{1,2}Johannes P van Dijk ^{1,4}Nele Vandenbussche ⁴Sigrid Pillen ^{1,4}Merel M van Gilst ^{1,4}Sebastiaan Overeem ^{1,4}

¹Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands; ²Philips Research, Eindhoven, the Netherlands; ³Sleep and Respiratory Care, Home Healthcare Solutions, Philips Austria GmbH, Vienna, Austria; ⁴Sleep Medicine Center Kempenhaeghe, Heeze, the Netherlands

*These authors contributed equally to this work

Purpose: There is great interest in unobtrusive long-term sleep measurements using wearable devices based on reflective photoplethysmography (PPG). Unfortunately, consumer devices are not validated in patient populations and therefore not suitable for clinical use. Several sleep staging algorithms have been developed and validated based on ECG-signals. However, translation from these techniques to data derived by wearable PPG is not trivial, and requires the differences between sensing modalities to be integrated in the algorithm, or having the model trained directly with data obtained with the target sensor. Either way, validation of PPG-based sleep staging algorithms requires a large dataset containing both gold standard measurements and PPG-sensor in the applicable clinical population. Here, we take these important steps towards unobtrusive, long-term sleep monitoring.

Methods: We developed and trained an algorithm based on wrist-worn PPG and accelerometry. The method was validated against reference polysomnography in an independent clinical population comprising 244 adults and 48 children (age: 3 to 82 years) with a wide variety of sleep disorders.

Results: The classifier achieved substantial agreement on four-class sleep staging with an average Cohen's kappa of 0.62 and accuracy of 76.4%. For children/adolescents, it achieved even higher agreement with an average kappa of 0.66 and accuracy of 77.9%. Performance was significantly higher in non-REM parasomnias (kappa = 0.69, accuracy = 80.1%) and significantly lower in REM parasomnias (kappa = 0.55, accuracy = 72.3%). A weak correlation was found between age and kappa ($\rho = -0.30$, $p < 0.001$) and age and accuracy ($\rho = -0.22$, $p < 0.001$).

Conclusion: This study shows the feasibility of automatic wearable sleep staging in patients with a broad variety of sleep disorders and a wide age range. Results demonstrate the potential for ambulatory long-term monitoring of clinical populations, which may improve diagnosis, estimation of severity and follow up in both sleep medicine and research.

Keywords: hypnogram, sleep staging, polysomnography, heart rate variability, wearable, pediatrics

Correspondence: Bernice M Wulterkens
Eindhoven University of Technology,
Department of Electrical Engineering,
Room Flux 7.104, PO BOX 513,
Eindhoven, 5600 MB, the Netherlands
Email b.m.wulterkens@tue.nl

Introduction

Sleep of adequate duration and quality is a central aspect of a healthy life. Many factors can contribute to insufficient or disturbed sleep, including a wide variety of sleep disorders. Objective assessment of sleep structure is an important part of the diagnostic work-up of patients with suspected sleep disorders. The current gold standard is polysomnography (PSG), using an array of body-worn sensors to assess

sleep, typically during a single night. Methods to perform long-term ambulatory monitoring of sleep would have important applications, such as more precise severity assessment, evaluation of night-to-night variability and follow-up of treatment. Moreover, the use of unobtrusive methods would provide additional practical benefits as it avoids interference with the sleep of the subjects and may be better accepted, especially in children.

Today, the consumer market is flooded with wearable devices promoting endless possibilities to trace a person's health by monitoring daily activities, energy expenditure and sleep. Typical consumer sleep trackers contain sensors measuring motion and cardiac activity. Advanced algorithms are then applied to perform sleep staging with the promise to distinguish wake, "light sleep", "deep sleep", and Rapid Eye Movement (REM) sleep.¹ Importantly, most manufacturers clearly state that their products are not intended for scientific or medical purposes and that care must be exercised when using and interpreting these data.^{2,3} Nevertheless, a growing number of people collect data from wearables and ask their physician to evaluate their self-measured sleep.⁴ On the other hand, the clinical use of wearable devices is gaining attention, given the potential advantages including unobtrusive sleep measurements collected in a patient's natural environment.¹ However, proper validation against gold standard PSG is scarcely available as of yet.⁴ Moreover, the limited validation efforts for popular devices are typically restricted to healthy participants, especially young adults with adequate sleep schedules and without sleep disorders.⁵

The combination of modern hardware capabilities and advanced machine learning methods for data analysis yields the promise of technology that can significantly improve the current diagnostic approach in sleep medicine. For example, the use of artificial intelligence to automatically score sleep stages based on PSG results in comparable and – in some cases – more consistent agreement than humans performing the same task.^{6,7} Automatic scoring of sleep stages based on signals that can be obtained from wearable devices still falls short in comparison to PSG based methods, although performance seems to be catching up.^{8–10} Most well-validated algorithms for what could be called "surrogate sleep staging" use heart rate variability (HRV) as an indicator of autonomic changes during sleep, which have been studied extensively over the last decades. Both the sympathetic nervous system (SNS) and parasympathetic nervous system (PNS), which are involved in the cardiac autonomic nervous system, are

coupled with circadian rhythm, the sleep-wake cycle, and ultradian processes such as NREM and REM sleep.¹¹ For example, in healthy subjects, autonomic cardiovascular regulation varies considerably per sleep stage. As NREM sleep progresses from light towards deeper sleep, there is an increase in cardiovagal drive and PNS activity and a reduction in cardiac and SNS activity. This results in a decrease in heart rate and increase in the respiratory mediation of HRV and becomes visible in the high-frequency band (HF) (0.15 to 0.4 Hz) of HRV. In contrast, autonomic activity is unstable during REM where PNS and SNS activity fluctuates producing abrupt changes in heart rate. The average heart rate and the power in the low-frequency band (LF) (0.04 to 0.15 Hz) of HRV is higher during REM than during NREM sleep, and there is a shift of the LF/HF ratio towards sympathetic dominance.¹¹ These changes in autonomic activity are often captured with different HRV features, which in turn can be derived from inter-beat intervals (IBIs) obtained by electrocardiography (ECG), while the vast majority of (consumer) sleep trackers use reflective photoplethysmography (PPG) to measure cardiac activity and, in addition, accelerometers to detect body movements. This leaves a gap between the available ECG-based algorithms and the practical application of such algorithms in the context of wearables. There are several factors contributing to this, such as the vulnerability of PPG to motion artefacts, which can impair the extraction of HRV features.¹² On the other hand, pulse transit time (PTT) may be affected by blood pressure, which is known to vary differently across sleep stages, thereby influencing differently PTT and in turn, PPG-derived beat intervals and consequently, HRV.^{13,14} These factors may be some of the reasons that ECG-derived algorithms cannot be used on PPG-based HRV without a significantly negative impact on performance.¹⁵ Accordingly, the translation from an ECG-based model to a clinically applicable PPG-based model is a non-trivial step and requires either the integration, in the model, of knowledge about the differences between the modalities, or alternatively, that the model is trained with data acquired with the target sensor, ie, PPG. Either approach requires a sufficient amount of data containing both gold standard PSG and PPG measurements from a clinical population. Here, we take this important step towards unobtrusive, long-term remote monitoring of sleep in the clinical setting. We developed an automatic sleep staging model using wrist-worn PPG signals and accelerometry, and trained and validated it on a unique,

large dataset containing data from a clinical population comprising children, adolescents and adults with a variety of sleep disorders.

Methods

Datasets

Data Acquisition and Sleep Stage Scoring

We used reference PSG and raw PPG and accelerometry to train and validate the automatic sleep staging model. PPG and accelerometer data were obtained with a CE-marked wrist-worn sensor (Philips Research, Eindhoven, the Netherlands), containing a three-axial accelerometer (sampling frequency: 128 Hz) and a PPG sensor with a light source consisting of two green light LEDs (sampling frequency: 32 Hz).¹⁹ Participants wore the device on their non-dominant wrist, with the sensor facing the skin on the dorsal side of the hand, above the ulnar styloid process.

To equalize the time base between PSG and PPG/accelerometer measurements, each recording was restricted to the period between lights off and lights on as determined during PSG. This period corresponds to the interval for which the PSG was scored with respect to sleep stages.

Each PSG recording was evaluated by a single scorer out of a team of seven certified, expert sleep technicians from the Sleep Medicine Center Kempenhaeghe (Heeze, the Netherlands). Kempenhaeghe is a third-line expert center for multidisciplinary sleep medicine. Institutional interrater agreement scores according to the American Academy of Sleep Medicine (AASM) assessment criteria are high, with an average agreement of 85.6% (range 83–88%). There were no systematic differences between recordings scored by different technicians for SOL, WASO or number of awakenings.²⁰

Sleep stages were scored in 30-second epochs according to the 2015 AASM criteria.²¹ To validate 4-class sleep staging, the ground-truth reference classes were obtained by combining stage N1 and N2 in a single “N1+N2” class, representing “light sleep”, while the remaining classes Wake, N3 (“deep sleep”) and REM were used without changes.

Training Dataset

Sleep Disordered Patients - The training dataset included all 422 sleep disordered patients (416 adults and 6 children/adolescents) available in the Sleep and Obstructive Sleep Apnea Measuring with Non-Invasive

Applications (SOMNIA) cohort by August 15th 2018.¹⁹ The SOMNIA dataset is created by Kempenhaeghe, and includes data from patients with a wide range of sleep disorders, including sleep-related breathing disorders, insomnia, sleep-related movement disorders and parasomnias. The primary sleep diagnosis was coded according to the International Classification of Sleep Disorders (ICSD) criteria.²² Where applicable, multiple sleep disorders could be entered. For analysis, subjects were grouped according to the ICSD main categories, as described by Fonseca et al.¹⁷

Healthy Sleepers - A total of 121 recordings from healthy adults were also included in the training dataset. From these, 81 recordings were obtained from two cohorts, the Night to Night (N2N) and Heart Health Study (HHS) datasets, collected by Philips Research during 2014 and 2015. Participants had no neurological, cardiovascular, psychiatric, pulmonary, endocrinological, or sleep disorders. In addition, people using sleep, antidepressant or cardiovascular medication, recreational drugs or excessive amounts of alcohol were excluded from the study, as well as pregnant women, shift workers and people who crossed more than two time zones in the last two months. The study protocols were described in detail by Fonseca et al.²³ The remaining 40 recordings were acquired from the HealthBed dataset, collected at the Sleep Medicine Center Kempenhaeghe, using the same protocol as the SOMNIA database.¹⁹ This set included all recordings that were available up to 15 October 2018. The HealthBed dataset comprises healthy adults without sleep disorders or other medical or psychiatric comorbidity.

The SOMNIA and HealthBed studies were reviewed by the medical ethical committee of the Maxima Medical Center (Eindhoven, the Netherlands. File no: N16.074 and W17.128). The Internal Committee of Biomedical Experiments of Philips Research approved the N2N and HHS studies. All studies met the ethical principles of the Declaration of Helsinki, the guidelines of Good Clinical Practice and the current legal requirements. The protocol for data analysis was approved by the Medical Ethical Committee of the Kempenhaeghe hospital and by the Internal Committee of Biomedical Experiments of Philips Research.

Hold-Out Validation Set

The model was validated on a separate hold-out validation set. No data of this set was used to train, tune, or otherwise

adapt the original model. This validation dataset comprised 244 recordings from adults and 48 recordings from children and adolescents with a wide variety of sleep disorders. Data was obtained from the SOMNIA dataset, as described above, and contained all available recordings collected between August 15th 2018 and January 8th 2020 for adults, and September 19th 2019 for children and adolescents.¹⁹

Algorithm

Previously, we developed a machine-learning model for automatic sleep staging based on long short-term memory (LSTM) recurrent neural networks. This model was initially developed, trained and validated on IBIs obtained from ECG data. All details on this model can be found in the relevant publications.^{16,17} Here we will provide a summary.

First, PPG signals were pre-processed by bandpass filtering between 0.3 and 5 Hz. Individual heartbeats were localized by detecting troughs (local minima) in the filtered waveform. The distance between consecutive heartbeats was calculated, and, using linear interpolation to 4 Hz, an IBI time series was built and used as input to the feature extraction step. A total of 132 HRV features were computed for each 30-second epoch in each PPG recording. HRV features were combined with a measure of gross body movements calculated as activity counts for each 30-second epoch based on the three-axial accelerometer signal.

The combined HRV and body movement feature set was used as input to a classifier with an input dense layer with 32 units, followed by a stack of three bi-directional LSTM layers with 64 units each and finally, two dense layers, the first with 32 units, and the last which outputs the posterior probability for each of four classes: Wake, N1+N2, N3 and REM for each 30-second epoch. The final classification is the class with the highest posterior probability for each epoch. The model used in this study has the same architecture as in our earlier work, but while in that study the HRV features used to train the classifier were extracted from ECG, in the current work, we trained it with a set comprising HRV features extracted from PPG and body movements obtained by an accelerometer, and used the corresponding manually scored sleep stages from PSG as ground-truth.¹⁷ The model was trained with the *RMSprop* optimization algorithm using categorical cross-entropy as a loss function.¹⁸ The training set, comprising data from both healthy sleepers and sleep

disordered patients, was split in a 75–25% ratio, with the largest portion used for model fitting. The second, smaller portion was used for early stopping to avoid overfitting, using as criteria to stop training for a lack of performance improvement for at least 50 consecutive training iterations. Using this criterion, the model was trained for a total of 1233 iterations.

Subsequently, we validated the model in a clinical sleep-disordered population using a separate hold-out validation set.

Analysis of Performance

Sleep stages classified by the model were compared to gold standard manual PSG scoring using measures described below, as proposed by De Zambotti et al.¹

Epoch-per-Epoch Agreement

Epoch-per-epoch agreement between the classified sleep stages by the algorithm and gold standard scored PSG sleep stages was evaluated using two quality metrics: Cohen's kappa coefficient of inter-rater agreement (or: κ) and accuracy. The first metric is an appropriate measure for quantifying the level of agreement for categorical data between two scorers, ie, in this work the classifications made by the algorithm against ground truth PSG. In addition, it is a more robust measure than accuracy, as κ takes into account the possibility of agreement occurring by chance. κ is usually interpreted with the following terms of agreement: $\kappa < 0$ "poor", $0 \leq \kappa \leq 0.20$ "slight", $0.20 < \kappa \leq 0.40$ "fair", $0.40 < \kappa \leq 0.60$ "moderate", $0.60 < \kappa \leq 0.80$ "substantial" and $0.80 < \kappa \leq 1.00$ "almost perfect".²⁴ Both κ and accuracy were computed for each recording, for 4-class sleep staging as described above. In addition, these metrics were computed for 3-class (merging N1+N2 and N3 in a single non-REM "NREM" class), and 2-class (merging N1+N2, N3 and REM in a single "Sleep" class) classification. For 2-class classification, we also calculated sensitivity, specificity and positive predictive value (PPV; all in respect to the detection of the positive class, ie, wake). A similar analysis was performed for the remaining classes (N1+N2, N3 and REM), considering each class (as positive) in comparison with the remaining, merged in a single (negative) class.

Confusion Matrix

A confusion matrix was plotted to further detail the proportion of PSG epochs correctly and incorrectly classified by the algorithm, with the advantage of also providing

information about the source of misclassification. The confusion matrix was obtained by aggregating the classifications and corresponding ground-truth of all epochs of all recordings of the hold-out validation set.

Influence of Demographic and Clinical Characteristics

The impact of demographic and clinical characteristics on 4-class epoch-per-epoch performance was assessed in various ways. Spearman's rank correlation was used to evaluate the effect of age and the influence of sex was assessed with the Wilcoxon rank-sum test. The effect of body mass index (BMI) on performance was tested using Spearman's rank correlations. This analysis was limited to adult subjects as BMI is not a metric that is appropriate to use in children. A detailed description of the influence of the apnea-hypopnea index (AHI) – an indicator of sleep disordered breathing severity – on performance can be found in the [Supplementary Data](#).

Performance in Relation to Sleep Disorder Diagnosis

In order to understand the classifier's behavior in the presence of different sleep disorders, we calculated κ and accuracy separately for each disorder category. The Wilcoxon rank-sum test was used to evaluate whether the performance differences with respect to the remaining participants were significant.

Sleep Statistics

As an intuitive indicator for the practical application of our model, we computed commonly used sleep statistics both on ground-truth PSG sleep stages and on the classified sleep stages by the algorithm. These included sleep onset latency (SOL), wake after sleep onset (WASO), total wake time (TWT), total sleep time (TST), sleep efficiency (SE) and time in N1+N2, N3 and REM. The average error (the

difference between the PSG-based and the HRV-based statistic), standard deviation (SD), 95% limits of agreement (LoA) corresponding to the mean difference $\pm 1.96 \times \text{SD}$, and root mean square error (RMSE) were computed. A positive mean difference value indicates that the statistic tends to be underestimated by the algorithm-based method in comparison with PSG. The presence of proportional bias in the estimated sleep statistics was assessed by calculating Spearman's rank correlation between, on one hand, the average of the sleep statistic estimated with PSG and the algorithm, and on the other hand, the difference of the sleep statistic estimated with PSG and the algorithm.

All data are represented as mean \pm SD unless otherwise stated. We used an alpha of 0.05 as significance level. All statistical analyses were conducted using Python (version 3.7).²⁵

Results

Cohort Description

Table 1 shows demographic information about the training and the validation dataset. Both datasets contained more men than women. The median age was lower for the validation set (46 years, IQR: 29–58 years) compared to the training dataset (52 years, IQR: 41–60 years), as the validation dataset comprised more children/adolescents. For adults, age did not significantly differ between the training and validation dataset (Wilcoxon rank-sum test $p = 0.097$). Median age for the adults in the training dataset was 52 years (IQR: 41–60 years) and 49 years (IQR: 35–61 years) in the validation dataset.

Table 2 indicates the total prevalence of each sleep disorder category in the validation dataset, together with the number of patients for whom each disorder was the

Table 1 Demographic Information for Participants in the Training and Validation Datasets

Parameter	Training dataset			Validation dataset		
	Total	Healthy sleepers	Sleep disordered patients	Total	Adults	Children/ adolescents
N (participants)	543	121	422	292	244	48
N Female [%]	225 [41.4]	67 [55.4]	158 [37.4]	106 [36.3]	86 [35.2]	20 [41.7]
Age [min., max.] (yrs)	49.4 \pm 15.4 [3, 86]	45.7 \pm 13.8 [18, 69]	50.5 \pm 15.7 [3, 86]	42.3 \pm 19.7 [3, 82]	48.4 \pm 15.3 [19, 82]	11.6 \pm 4.4 [3, 17]
BMI (kg/m ²)	27.0 \pm 5.0	25.2 \pm 3.7	27.5 \pm 5.23	–	27.2 \pm 5.0	–

Table 2 Prevalence of Sleep Disorders in the Validation Dataset

	Adults		Children/Adolescents	
Group	Total Prevalence	Single Primary Disorder	Total Prevalence	Single Primary Disorder
Sleep disordered breathing	114	86	15	15
Insomnia	71	40	11	8
Movement disorder	35	14	2	2
Behavioral sleep disorder	22	9	4	3
Non-REM parasomnia	15	12	6	4
REM parasomnia	17	6	0	0
Circadian disorder*	-	-	3	3
Other	34	17	2	1
None	0	-	8	-

Notes: The number of participants with the respective diagnoses are shown as a total; as well as the number of participants in whom the respective diagnosis was the single primary sleep disorder. *Circadian disorder was evaluated as a separate group for children/adolescents. For adults, the circadian disorder group was incorporated in the category "other" as it contained less than 10 participants.

single primary disorder. For adults, the most often occurring combination of sleep disorders were sleep disordered breathing and insomnia (in six patients), and insomnia in combination with a sleep related movement disorder, also occurring in six patients. In the children/adolescents group, three participants had insomnia and a comorbid sleep disorder, including behavior-related sleep disorder, non-REM parasomnia or sleep-related bruxism.

Sleep Staging Performance

Epoch-per-Epoch Agreement

Table 3 shows the agreement between the algorithm-based sleep stage classifications and ground-truth PSG sleep stages in the validation dataset for the different sleep staging tasks.

As expected, the most difficult task (4-class sleep staging) showed the lowest performance, which nonetheless was substantial (κ 0.62 ± 0.12 , accuracy 76.4 ± 7.3). Overall, 3-class sleep staging showed the best performance. When assessing binary classification, ie, one sleep stage versus the rest, the performance was best for REM and worst for N1+N2.

Confusion Matrix

Table 4 shows the confusion matrix for the sleep stage classifications in all epochs of all recordings (a grand total of 298,219 epochs). Sleep stage N1+N2 was the most prevalent sleep stage. Most confusion in the classification of Wake, deep sleep (N3) and REM sleep occurred with this N1+N2 class. The two least prevalent classes (REM

Table 3 Overall Epoch-per-Epoch Agreement for Both Adults and Children/Adolescents

Task	κ (-)	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)
Wake/N1+N2/N3/REM	0.62 ± 0.12	76.4 ± 7.3	n/a	n/a	n/a
Wake/NREM/REM	0.68 ± 0.11	85.2 ± 5.8	n/a	n/a	n/a
Wake (vs Sleep) ^a	0.66 ± 0.14	91.5 ± 5.4	73.1 ± 16.8	94.6 ± 5.9	74.3 ± 16.6
N1+N2 ^a	0.54 ± 0.13	77.5 ± 6.7	78.0 ± 9.0	76.7 ± 11.6	78.8 ± 11.3
N3 ^a	0.60 ± 0.22	90.8 ± 4.9	69.5 ± 24.3	94.8 ± 4.9	69.1 ± 24.8
REM ^a	0.69 ± 0.18	93.0 ± 3.9	78.2 ± 19.6	95.2 ± 3.7	71.8 ± 16.1

Note: ^aBinary classification tasks were assessed by a one versus the rest strategy, where one single class (Wake, N1+N2, N3 or REM) was considered as the "positive" class and the remaining classes were aggregated in a single "negative" class.

Abbreviation: PPV, positive predictive value.

Table 4 Confusion Matrix for Sleep Stage Classification in All Epochs of All Recordings (N = 298,219)

Pred → Ref ↓	Wake	N1+N2	N3	REM	Prev. (-, (%))	Sens. (%)	κ (-)
Wake	41,962 (14.1%/75.7%)	11,900 (4.0%/21.5%)	125 (0.04%/0.2%)	1468 (0.5%/2.6%)	55,455 (18.6)	75.7	0.72
N1+N2	10,595 (3.6%/6.8%)	122,603 (41.1%/78.2%)	12,965 (4.3%/8.3%)	10,570 (3.5%/6.7%)	156,733 (52.6)	78.2	0.55
N3	310 (0.1%/0.7%)	13,480 (4.5%/30.0%)	30,795 (10.3%/68.8%)	279 (0.09%/0.6%)	201,907 (15.0)	68.6	0.64
REM	837 (0.3%/2.0%)	7560 (2.5%/18.4%)	256 (0.09%/0.6%)	32,514 (10.9%/79.0%)	41,167 (13.8)	79.0	0.72
PPV (%)	78.1	78.8	69.8	72.5			

Notes: Each entry in the confusion matrix indicates the number of epochs. Between parentheses, the percentage relative to the total number of epochs of all classes is listed, followed by the percentage relative to the total number of epochs with the corresponding reference sleep stage for that row.

Abbreviations: Prev., prevalence; Sens., sensitivity; PPV, positive predictive value.

and N3) showed the best and the second worst overall agreement, respectively, suggesting that the occurrence of a class is not associated with performance.

Influence of Demographic and Clinical Characteristics

Spearman's rank correlation coefficients showed a weak association between age and κ ($\rho = -0.30$, $p < 0.001$) and age and accuracy ($\rho = -0.22$, $p < 0.001$), indicating a slight decrease in performance with increasing age. This was also suggested by a higher performance in children/adolescents compared to adults. A detailed description of the performance in children/adolescents can be found in the Supplementary data [Table S1](#). A significant difference in performance was found between sexes for κ (average κ : female = 0.63, male = 0.61, $p < 0.05$), but not for accuracy (average accuracy: female =

0.77, male = 0.76, $p = 0.27$). For adults, no significant correlations were found between BMI and κ or accuracy.

Performance in Relation to Sleep Disorder Diagnosis

[Table 5](#) shows the 4-class sleep stage classification performance per diagnosis category. The performance for patients with sleep disordered breathing was lower with respect to κ but not accuracy, compared to patients without this specific diagnosis ($\kappa = 0.60 \pm 0.12$ vs $\kappa = 0.63 \pm 0.11$, $p = 0.024$). The performance for patients with a diagnosis of non-REM parasomnia was significantly higher (both in κ as in accuracy) than for patients without that disorder. In contrast, for REM parasomnias, performance was lower than for patients without that condition. A detailed look at the performance for non-REM parasomnias and REM parasomnias can be found in the Supplementary data [Tables S2](#)

Table 5 Performance for 4-Class Sleep Staging in Diagnostic Subgroups

Condition ^a	N	κ (-)			Accuracy (%)		
		Mean \pm SD	Median	P-value ^b	Mean \pm SD	Median	P-value ^c
Sleep disordered breathing	129	0.60 \pm 0.12	0.61	0.024*	75.57 \pm 7.74	77.53	0.15
Insomnia	82	0.63 \pm 0.11	0.63	0.43	77.16 \pm 6.69	77.15	0.64
Movement disorder	37	0.61 \pm 0.12	0.61	0.60	76.38 \pm 8.24	78.38	0.89
Behavioral	26	0.62 \pm 0.10	0.62	0.81	76.97 \pm 5.80	77.16	0.96
Non-REM parasomnia	21	0.69 \pm 0.07	0.70	0.0023**	80.06 \pm 4.32	80.73	0.012*
REM parasomnia	17	0.55 \pm 0.11	0.56	0.013*	72.30 \pm 7.56	72.39	0.013*

Notes: ^aSubgroup of patients for whom the primary diagnosis includes that disorder. ^{b,c}Wilcoxon rank-sum test of differences in κ and accuracy, respectively, between the subgroup of patients with that disorder and without. Bold values are statistically significant. ** $p < 0.01$, * $p < 0.05$.

and S3. For each of the remaining groups, no significant differences were found between patients with and without the respective disorder.

Sleep Statistics

Figure 1 illustrates a comparison between illustrative reference PSG hypnograms and algorithm-based hypnograms

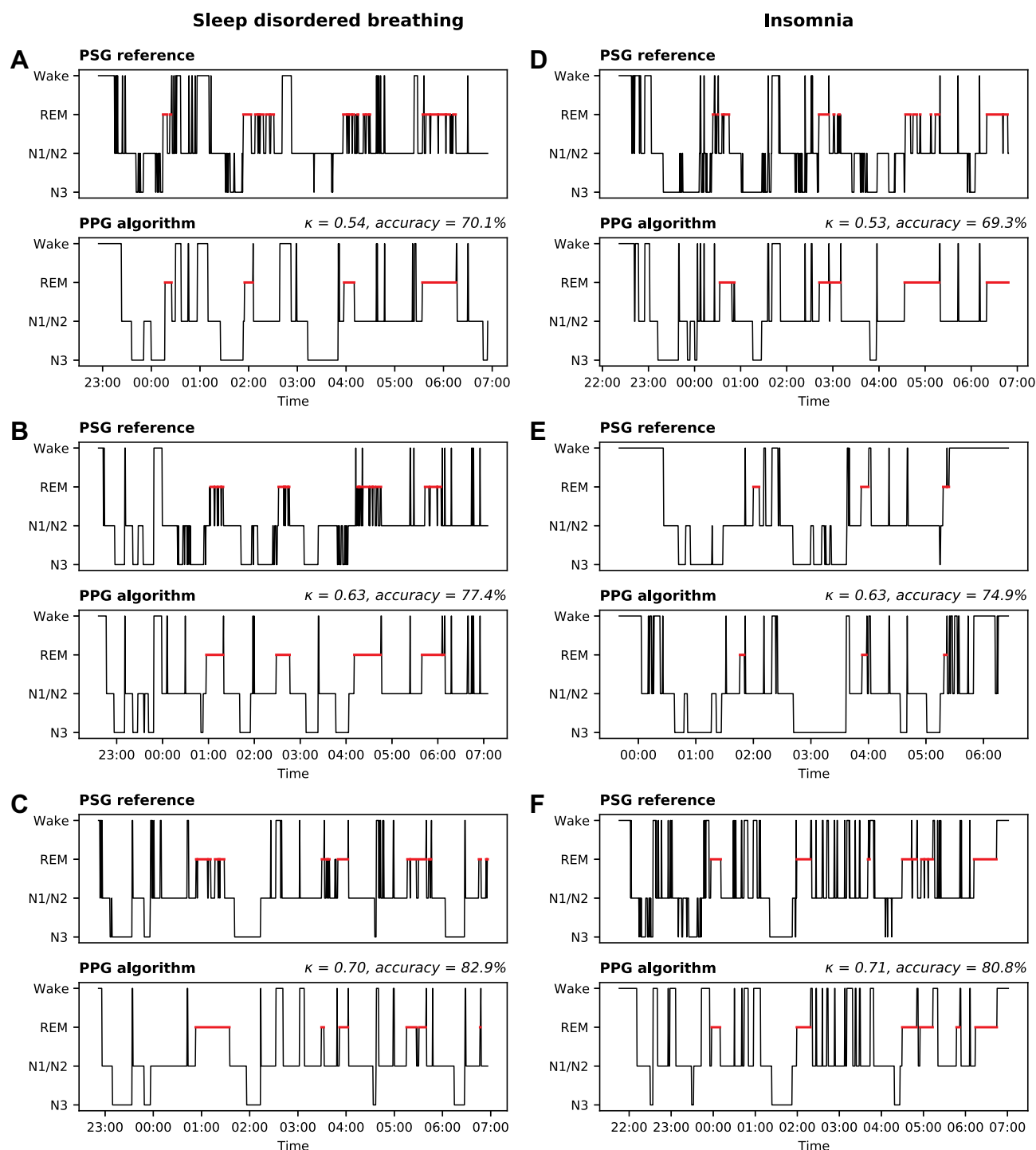


Figure 1 Representative hypnograms of three patients with a single primary diagnosis of sleep disordered breathing (left panel: patient (A–C)) and three patients with a single primary diagnosis of insomnia (right panel: patient (D–F)). Hypnograms are shown based on the PSG reference (top) and the PPG/accelerometer algorithm (bottom). Hypnograms were taken from the 25, 50 and 75 percentiles of overall kappa with the lowest performance on top. REM sleep is marked with a red line. Some clinical aspects are relevant to mention. Patient (B) was diagnosed with very mild, but treatment responsive obstructive sleep apnea, but also had parasomnia complaints. Patient (D and E) had both sleep misperception and were thus diagnosed with paradoxical insomnia. Patient (E) was also diagnosed with a delayed sleep phase, explaining the occurrence of N3-sleep later in the night. Patient (F) was diagnosed with insomnia, receiving quetiapine at the time of PSG with good results.

for three patients with sleep disordered breathing (A, B, and C) and three patients with insomnia (D, E, and F), based on the values of 25, 50 and 75 percentiles of overall kappa ($\kappa = 0.54$, $\kappa = 0.62$, $\kappa = 0.70$). The hypnograms show that not only overall performance is adequate, but that the overall sleep architecture is captured as well, allowing clinical interpretation. From the examples, it is clear that the detection of wake is generally accurate, which is important in the assessment of sleep quality in general and sleep disorders in particular, for example in patients with insomnia. A detailed description of the performance of our model in insomnia is further provided in the Supplementary data [Table S4](#). [Figure 1](#) also illustrates that the algorithm has some difficulties with fragmentation of sleep stages in some cases.

[Table 6](#) compares relevant overall sleep statistics estimated with PPG/accelerometer data versus PSG for adults and children/adolescents. Overall, the bias for all sleep statistics was relatively small, although dispersion was present between subjects. The estimation of sleep efficiency with the algorithm was almost identical compared to PSG; bias was less than 1% with a SD below 10%. RMSE was smallest for REM sleep, which was also identified as the best performing class according to epoch-per-epoch agreement.

Spearman's rank correlation showed a weak positive proportional bias indicating that the degree of the

underestimation increased as wake after sleep onset increased ($\rho = 0.19$, $p < 0.05$). A weak negative proportional bias was found for sleep onset latency and REM, indicating that the degree of overestimation increased as sleep latency increased ($\rho = -0.22$, $p < 0.001$) and time in REM sleep increased ($\rho = -0.17$, $p < 0.001$). The Bland-Altman plots for sleep onset latency, wake after sleep onset, total sleep time and sleep efficiency are shown in the Supplementary data [Figure S1](#).

Discussion

We developed an automatic sleep staging algorithm using PPG and accelerometry data obtained from a wrist-worn device. To our knowledge, this is the first study that evaluated the performance of a PPG-based sleep staging algorithm in such a large clinical population comprising adults, adolescents and children. Patients had a wide variety of sleep disorders, and many of them had multiple primary sleep disorders. Even for four-class sleep staging, the algorithm achieved substantial agreement with gold standard PSG. This agreement is in the same κ class as human interscorer agreement for five-class sleep staging.²⁶ Predominant confusions of our wearable method pertained between N1+N2 versus wake, N3 and REM. The discrepancy between N3 and N2 was also highest when comparing two human expert scorers in another study.²⁶

Table 6 Sleep Statistics for Both Adults and Children/Adolescents

Parameter	PSG		PSG – Sleep Statistic Calculated by the Algorithm		
	Mean \pm SD	Range [min., max.]	Mean Error \pm SD	95% LoA	RMSE
SOL (min)	18.68 \pm 22.46	[0.00, 221.00]	-5.28 \pm 24.83	[-53.94, 43.38]	25.34
WASO (min)	72.33 \pm 65.55	[4.50, 391.00]	9.80 \pm 39.45	[-67.52, 87.13]	40.59
TWT (min)	94.96 \pm 76.17	[5.00, 492.00]	2.93 \pm 33.98	[-63.67, 69.53]	34.05
TST (min)	415.00 \pm 85.32	[102.00, 651.50]	-3.99 \pm 34.08	[-70.78, 62.81]	34.26
SE (%)	81.25 \pm 14.63	[19.01, 99.01]	-0.70 \pm 6.63	[-13.70, 12.30]	6.66
Time in N1+N2 (min)	267.68 \pm 60.99	[47.00, 434.00]	1.24 \pm 49.95	[-96.66, 99.15]	49.88
Time in N1+N2 (%)	65.02 \pm 10.63	[33.33, 100.00]	0.59 \pm 10.65	[-20.29, 21.47]	10.65
Time in N3 (min)	76.82 \pm 40.69	[0.00, 237.50]	1.19 \pm 40.25	[-77.70, 80.08]	40.20
Time in N3 (%)	18.49 \pm 9.36	[0.00, 66.67]	-0.62 \pm 9.56	[-18.12, 19.35]	9.56
Time in REM (min)	70.49 \pm 30.77	[0.00, 164.50]	-6.42 \pm 22.92	[-51.36, 38.30]	23.77
Time in REM (%)	16.49 \pm 5.96	[0.00, 34.63]	-1.21 \pm 5.80	[-12.58, 10.16]	5.92

Abbreviations: PSG, polysomnography; SD, standard deviation; min., minimum; max., maximum; LoA, limits of agreement; RMSE, root mean square error; SOL, sleep onset latency; WASO, wake after sleep onset; TWT, total wake time; TST, total sleep time; SE, sleep efficiency; min, minutes.

Importantly, the effect of age on sleep staging performance was limited, indicating that the algorithm is robust across all age categories.

We investigated the clinical utility of our method by comparing sleep measures estimated by the algorithm with those derived from PSG. Only a small average bias was found for each sleep statistic, albeit with variations between subjects. Our wearable method can thus provide clinically relevant information about a patient's sleep pattern, especially when visualized as a hypnogram. A difference in performance was found between men and women, but not of an extent that would imply consequences for clinical use. Importantly, performance and estimation of sleep statistics were even better for children and adolescents, despite the limited number of children/adolescents in the training dataset. The reason for this is not fully clear and several factors may contribute. For example, one could speculate that children have less sleep stage transitions. Our algorithm shows sometimes difficulties with the detection of high sleep fragmentation resulting in a better performance on patients with lower fragmentation. In addition, autonomic expression, and in particular parasympathetic activity, is stronger in children, as it is known that parasympathetic tone decreases with increasing age. We could assume that the performance of the algorithm is better in younger subjects with higher PNS activity, since it becomes easier to link high parasympathetic tone with sleep stage N3, regardless of how it was trained. Finally, the performance of our algorithm was lowest for patients with REM parasomnias and this sleep disorder was not present in the children/adolescents group, which could have contributed to the higher performance in this group.

Moreover, commercial wearables have mainly shown less reliable results in this population so far.^{27,28} As sleep disturbances are relatively common during childhood and PSG with its large number of body-worn sensors is often not well tolerated by children, a less obtrusive alternative for reliable sleep monitoring is highly needed.^{29,30} For example, infants, or children with developmental delays are unable to report their sleep patterns. In addition, parents may become less involved with bedtime routines over time and may not be aware of their child's sleep onset latency or nighttime awakenings.^{31,32}

Our results achieved in this clinical population demonstrate an improved performance compared to literature, even though studies reported so far were mainly performed in healthy populations. An overview of recent literature on

automatic sleep staging using PPG signals, similar to our work, can be found in the Supplementary data [Table S5](#). These studies achieved a kappa between 0.38 and 0.54, and accuracies between 60% and 69% for 4-class classification.^{33–35} Our method achieved a higher performance on 4-class classification with an average kappa of 0.62 and an accuracy of 76.4%, illustrating substantial improvement compared to the reported methods. A small note should be made that direct comparison of our results with the results reported in literature is slightly difficult due to the different study populations. In order to evaluate the performance of the model in healthy participants, it would be necessary to collect PPG and accelerometer data in this population.

Importantly, our algorithm shows relatively high agreement for 2-class Wake/Sleep detection, which is a great advantage over actigraphy. Actigraphy is currently considered the most important clinical tool to obtain long-term, objective data on sleep-wake structure at home, despite its major limitation: its low ability to accurately detect wake.³² In the widely available consumer wearables, the sensitivity to estimate wake is catching up in healthy subjects, up to 83% in adolescents.^{36–38} However, the performance in sleep disordered patients still falls short.^{39,40} This highlights the importance for careful interpretation of data collected by these consumer devices. Users may be convinced of having a sleep disorder based on the feedback provided by the sleep tracker, even when this is not the case. This condition was recently coined orthosomnia and sleep clinics are increasingly confronted with this phenomenon.⁴¹ An opposite, potentially perverse effect of an insufficient performance is that sleep trackers may fail to detect the presence of clinically relevant sleep disruptions, falsely reassuring the user while a sleep disorder remains undiagnosed. An overview of recently published research on the ability to detect wake using actigraphy and consumer wearables can be found in the Supplementary data [Table S5](#). These data show that our method achieved substantial improvements in performance compared to those available so far.

Our algorithm had most difficulties with classifying sleep stage N1+N2 and N3 relative to all sleep stages. This phenomenon is quite similar to (dis)agreement between human scoring in the assessment of slow waves using electroencephalogram (EEG) signals.⁴² When using a classification model based on HRV, this problem is slightly different, but still comparable. It is difficult to establish from which point onward the parasympathetic

tone, used as an autonomic hallmark for sleep stage N3, is strong enough to change the classification from sleep stage N2 to N3. These “continuous” changes in the autonomic spectrum remain an obvious limitation of the technology, at least when attempting to provide a surrogate to classical epoch-by-epoch PSG sleep staging.

Another important finding is that the algorithm sometimes shows difficulties with the detection of sleep fragmentation, despite the overall accurate representation of standard sleep architecture parameters. Short non-REM intrusions in REM sleep periods, for instance, are only occasionally detected by the algorithm (see illustrative fragmented REM sleep periods in [Figure 1](#)). Note that, according to the AASM scoring rules, a mixture of rapid eye-movements and spindles and/or K complexes should be scored as either stage REM or stage N2 depending on the actual occurrence of these events in the respective 30-second epochs.²¹ However, the algorithm identifies these sleep stages as REM sleep. Frequent stage shifts between N3 and N2 (see, eg, N3 sleep in patient B in [Figure 1](#)) are typically seen in periods where just about 20% of an epoch consists of slow wave activity as the threshold to score N3. The algorithm seems to be less sensitive to these borderline cases.

Conclusion

In summary, this study shows the ability of automatic sleep staging using PPG and accelerometer measurements obtained with a wrist-worn device, in children, adolescents and adults with a wide variety of sleep disorders. The results demonstrate the maturity of this technique and the opportunities for remote, clinical, long-term sleep monitoring. This may yield a significant change in the clinical approach to diagnosis and follow up of sleep disorders. In addition, it will allow a fundamentally new view on night-to-night variability of sleep as a basis for new pathophysiological insights.

Abbreviations

AASM, American Academy of Sleep Medicine; AHI, apnea-hypopnea index; BMI, body mass index; ECG, electrocardiography; EEG, electroencephalogram; HHS, Heart Health Study; HRV, Heart rate variability; IBIs, inter-beat intervals; ICSD, International Classification of Sleep Disorders; IQR, inter quartile range; LoA, limits of agreement; LSTM, long short-term memory; N2N, Night to Night; OSA, obstructive sleep apnea; PNS,

parasympathetic nervous system; PPG, photoplethysmography; PPV, positive predictive value; PSG, polysomnography; REM, Rapid Eye Movement; RMSE, root mean square error; SD, standard deviation; SE, sleep efficiency; SNS, sympathetic nervous system; SOL, sleep onset latency; SOMNIA, Sleep and Obstructive Sleep Apnea Measuring with Non-Invasive Applications; TST, total sleep time; TWT, total wake time; WASO, wake after sleep onset.

Ethical Approval

Data were obtained from four different datasets. The SOMNIA and HealthBed studies were reviewed by the medical ethical committee of the Maxima Medical Center (Eindhoven, the Netherlands. File no: N16.074 and W17.128). The Internal Committee of Biomedical Experiments of Philips Research approved the N2N and HHS studies. All studies met the ethical principles of the Declaration of Helsinki, the guidelines of Good Clinical Practice and the current legal requirements. The protocol for data analysis was approved by the Medical Ethical Committee of the Kempenhaeghe hospital and by the Internal Committee of Biomedical Experiments of Philips Research.

Acknowledgments

We thank Mustafa Radha and Arnaud Moreau for contributions for previous versions of the algorithm, and Roy Krijn and Bertram Hoondert for help with data acquisition.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work. Merel M van Gilst and Sebastiaan share senior authorship.

Funding

This work was performed within the IMPULS framework of the Eindhoven MedTech Innovation Center (e/MTIC, incorporating Eindhoven University of Technology, Philips Research, and Sleep Medicine Center Kempenhaeghe), including a PPS-supplement from the

Dutch Ministry of Economic Affairs and Climate Policy. Additional support by STW/IWT in the context of the OSA+ project (No. 14619).

Disclosure

At the time of writing, PF, MR, AC, PA and XL were employed and/or affiliated with Royal Philips, a commercial company and manufacturer of consumer and medical electronic devices, commercializing products in the area of sleep diagnostics and sleep therapy. Philips had no role in the study design, decision to publish or preparation of the manuscript. SO received an unrestricted research grant from UCB Pharma and participated in advisory boards for UCB Pharma, Jazz Pharmaceuticals and Bioprojet, all paid to institution and all unrelated to the present work. Dr Pedro Fonseca reports personal fees from Philips Research during the conduct of the study; personal fees from Philips Research, outside the submitted work. Mr Marco Ross is an employee of Philips Austria GmbH, during the conduct of the study. Mr Andreas Cerny is an employee of Philips Austria GmbH, during the conduct of the study. Dr Peter Anderer is part time employee of Philips Austria, during the conduct of the study. Dr Xi Long is an employee of Royal Philips, during the conduct of the study. Dr Sigrid Pillen reports personal fees for consulting for Conect4Children, and for presentations and education in the field of pediatric sleep medicine, outside the submitted work. The other authors have no conflicts of interest.

References

- De Zambotti M, Cellini N, Goldstone A, Colrain IM, Baker FC. Wearable sleep technology in clinical and research settings. *Med Sci Sports Exerc.* 2019;51(7):1538–1557. doi:10.1249/MSS.0000000000001947
- Garmin. Accuracy; 2020. Available from: <https://www.garmin.com/en-US/legal/atdisclaimer/>. Accessed October 28, 2020.
- Fitbit. Fitbit important safety and product information; 2020. Available from: <https://www.fitbit.com/global/us/legal/safety-instructions>. Accessed October 28, 2020.
- Seema K, Deak Maryann C, Dominic G, et al. Consumer sleep technology: an American Academy of Sleep Medicine position statement. *J Clin Sleep Med.* 2018;14(5):877–880. doi:10.5664/jcsm.7128
- Baron KG, Duffecy J, Berendsen MA, Cheung Mason I, Lattie EG, Manalo NC. Feeling validated yet? A scoping review of the use of consumer-targeted wearable and mobile technology to measure and improve sleep. *Sleep Med Rev.* 2018;40:151–159. doi:10.1016/j.smrv.2017.12.002
- Zhang L, Fabbri D, Upender R, Kent D. Automated sleep stage scoring of the Sleep Heart Health Study using deep neural networks. *Sleep.* 2019;42(11). doi:10.1093/sleep/zsz159
- Stephansen JB, Olesen AN, Olsen M, et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat Commun.* 2018;9(1):5229. doi:10.1038/s41467-018-07229-3
- Lee XK, Chee NIYN, Ong JL, et al. Validation of a consumer sleep wearable device with actigraphy and polysomnography in adolescents across sleep opportunity manipulations. *J Clin Sleep Med.* 2019;15(09):1337–1346. doi:10.5664/jcsm.7932
- de Zambotti M, Goldstone A, Claudatos S, Colrain IM, Baker FC. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. *Chronobiol Int.* 2018;35(4):465–476. doi:10.1080/07420528.2017.1413578
- de Zambotti M, Rosas L, Colrain IM, Baker FC. The sleep of the ring: comparison of the Oura sleep tracker against polysomnography. *Behav Sleep Med.* 2019;17(2):124–136. doi:10.1080/15402002.2017.1300587
- Lanfranchi PA, Pépin JL, Somers VK. Cardiovascular physiology: autonomic control in health and in sleep disorders. In: *Principles and Practice of Sleep Medicine*. 6th edition. Elsevier Inc.; 2016:142–154.
- Pietilä J, Mehra S, Tolonen J, et al. Evaluation of the accuracy and reliability for photoplethysmography based heart rate and beat-to-beat detection during daily activities. In: Eskola H, Väisänen O, Viik J, Hyttinen J, editors. *EMBE & NBC 2017. IFMBE Proceedings*. Springer; 2018:145–148. doi:10.1007/978-981-10-5122-7_37
- Asmar R, Benetos A, Topouchian J, et al. Assessment of arterial distensibility by automatic pulse wave velocity measurement. *Hypertension.* 1995;26(3):485–490. doi:10.1161/01.HYP.26.3.485
- Tripathi A, Obata Y, Ruzankin P, et al. A pulse wave velocity based method to assess the mean arterial blood pressure limits of autoregulation in peripheral arteries. *Front Physiol.* 2017;8:855. doi:10.3389/fphys.2017.00855
- van Gilst MM, Wulterkens BM, Fonseca P, et al. Direct application of an ECG-based sleep staging algorithm on reflective photoplethysmography data decreases performance. *BMC Res Notes.* 2020.
- Radha M, Fonseca P, Moreau A, et al. Sleep stage classification from heart-rate variability using long short-term memory neural networks. *Sci Rep.* 2019;9(1):1–11. doi:10.1038/s41598-019-49703-y
- Fonseca P, Gilst MM, Radha M, et al. Automatic sleep staging using heart rate variability, body movements and recurrent neural networks in a sleep disordered population. *Sleep.* 2020;43. doi:10.1093/sleep/zsaa048
- Hinton G. Neural networks for machine learning online course. Coursera; 2021. Available from: <https://www.coursera.org/learn/neural-networks-deep-learning>. Accessed March 26, 2021.
- Gilst MM, Dijk JP, Krijn R, et al. Protocol of the SOMNIA project: an observational study to create a neurophysiological database for advanced clinical sleep monitoring. *BMJ Open.* 2019;9(11). doi:10.1136/bmjopen-2019-030996
- Hermans LWA, Van Gilst MM, Regis M, et al. Modeling sleep onset misperception in insomnia. *Sleep.* 2020;43(8). doi:10.1093/sleep/zsaa014
- Berry RB, Brooks R, Gamaldo CE, Harding SM, Marcus C, Vaughn BV. The AASM manual for the scoring of sleep and associated events. *Rules Terminol Tech Specif Darien Ill Am Acad Sleep Med.* 2015;176:2015.
- Medicine AA of S. International classification of sleep disorders. *Diagn Coding Man.* 2005;51–55.
- Fonseca P, Weysen T, Goelema MS, et al. Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle-aged adults. *Sleep.* 2017;40(7). doi:10.1093/sleep/zsx097
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159–174. doi:10.2307/2529310
- Python Software Foundation. Python language reference, version 3.7; 2021. Available from: <http://www.python.org>. Accessed May 11, 2021.
- Danker-Hopfe H, Anderer P, Zeitlhofer J, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *J Sleep Res.* 2009;18(1):74–84. doi:10.1111/j.1365-2869.2008.00700.x
- Meltzer LJ, Hiruma LS, Avis K, Montgomery-Downs H, Valentin J. Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. *Sleep.* 2015;38(8):1323–1330. doi:10.5665/sleep.4918

28. de Zambotti M, Baker FC, Colrain IM. Validation of sleep-tracking technology compared with polysomnography in adolescents. *Sleep*. 2015;38(9):1461–1468. doi:10.5665/sleep.4990
29. Esposito S, Laino D, D'Alonzo R, et al. Pediatric sleep disturbances and treatment with melatonin. *J Transl Med*. 2019;17(1):77. doi:10.1186/s12967-019-1835-1
30. Malow BA, Marzec ML, McGrew SG, Wang L, Henderson LM, Stone WL. Characterizing sleep in children with autism spectrum disorders: a multidimensional approach. *Sleep*. 2006;29(12):1563–1571. doi:10.1093/sleep/29.12.1563
31. Acebo C, Sadeh A, Seifer R, Tzischinsky O, Hafer A, Carskadon MA. Sleep/wake patterns derived from activity monitoring and maternal report for healthy 1- to 5-year-old children. *Sleep*. 2005;28(12):1568–1577. doi:10.1093/sleep/28.12.1568
32. Meltzer LJ, Avis KT, Biggs S, Reynolds AC, Crabtree VM, Bevens KB. The Children's Report of Sleep Patterns (CRSP): a self-report measure of sleep for school-aged children. *J Clin Sleep Med*. 2013;9(3):235–245. doi:10.5664/jcsm.2486
33. Beattie Z, Oyang Y, Statan A, et al. Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals. *Physiol Meas*. 2017;38(11):1968–1979. doi:10.1088/1361-6579/aa9047
34. Korkalainen H, Aakko J, Duce B, et al. Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea. *Sleep*. 2020. doi:10.1093/sleep/zsaa098
35. Molkari M, Tenhunen M, Tarniceriu A, Vehkaoja A, Himanen S-L, Räsänen E. Non-linear heart rate variability measures in sleep stage analysis with photoplethysmography. *Comput Cardiol*. 2019. doi:10.22489/CinC.2019.287
36. Godino JG, Wing D, Zambotti M, et al. Performance of a commercial multi-sensor wearable (Fitbit Charge HR) in measuring physical activity and sleep in healthy children. *PLoS One*. 2020;15(9):e0237719. doi:10.1371/journal.pone.0237719
37. Pesonen A-K, Kuula L. The validity of a new consumer-targeted wrist device in sleep measurement: an overnight comparison against polysomnography in children and adolescents. *J Clin Sleep Med*. 2018;14(4):585–591. doi:10.5664/jcsm.7050
38. Chinoy ED, Cuellar JA, Huwa KE, et al. Performance of seven consumer sleep-tracking devices compared with polysomnography. *Sleep*. 2021. doi:10.1093/sleep/zsaa291
39. Kahawage P, Jumabhoy R, Hamill K, Zambotti M, Drummond SPA. Validity, potential clinical utility, and comparison of consumer and research-grade activity trackers in Insomnia Disorder I: in-lab validation against polysomnography. *J Sleep Res*. 2020;29(1). doi:10.1111/jsr.12931
40. Moreno-Pino F, Porras-Segovia A, López-Esteban P, Artés A, Baca-García E. Validation of Fitbit Charge 2 and Fitbit Alta HR against polysomnography for assessing sleep in adults with obstructive sleep apnea. *J Clin Sleep Med*. 2019;15(11):1645–1653. doi:10.5664/jcsm.8032
41. Glazer BK, Sabra A, Nancy J, Natalie M, Rebecca M. Orthosomnia: are some patients taking the quantified self too far? *J Clin Sleep Med*. 2017;13(2):351–354. doi:10.5664/jcsm.6472
42. Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. *J Clin Sleep Med*. 2013;9(1):81–87. doi:10.5664/jcsm.2350

Nature and Science of Sleep

Publish your work in this journal

Nature and Science of Sleep is an international, peer-reviewed, open access journal covering all aspects of sleep science and sleep medicine, including the neurophysiology and functions of sleep, the genetics of sleep, sleep and society, biological rhythms, dreaming, sleep disorders and therapy, and strategies to optimize healthy sleep.

Submit your manuscript here: <https://www.dovepress.com/nature-and-science-of-sleep-journal>

Dovepress

The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.